AI

MDPI

Article

CrossRate: A Label-Free Measure for Sentiment Analysis in Visual Emotion Recognition

Gintautas Dzemyda and Modestas Motiejauskas *

Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University, Akademijos Street 4, LT-08412 Vilnius, Lithuania; gintautas.dzemyda@mif.vu.lt
* Correspondence: modestas.motiejauskas@mif.vu.lt

Abstract

Images may evoke different emotional responses depending on their content, style, and viewer interpretation. This is particularly important when evaluating works of art, architectural designs, or interior design choices. Visual emotion recognition (VER) models are commonly evaluated using supervised classification metrics such as accuracy, precision, recall, and macro-F1, together with general uncertainty indicators such as entropy, maximum softmax probability (MSP), and top-1/top-2 margin. However, these measures do not indicate whether the model's strongest competing-emotion predictions remain within the same sentiment group or cross the positive–negative sentiment boundary. This paper proposes the top-2 cross-sentiment rate (CrossRate), a label-free measure for analyzing sentiment-level ambiguity in VER models. CrossRate measures the proportion of samples for which the top-1 and top-2 predicted emotion classes belong to opposite sentiment groups. The measure is evaluated on VER datasets using both standard classification metrics and uncertainty indicators. Experiments on EmoSet-118K show that varying the model's parameters reduces CrossRate from $(22.15 \pm 0.45)\%$ to $(7.81 \pm 0.62)\%$ and increases accuracy from $(79.13 \pm 0.16)\%$ to $(80.10 \pm 0.15)\%$. These changes are not fully reflected by entropy, MSP, or margin, indicating that CrossRate captures a complementary aspect of sentiment-level prediction behavior. The WikiArt case study further demonstrates that CrossRate can be applied when ground-truth emotion labels are unavailable. The proposed measure is applicable to any VER model whose predicted emotion classes can be mapped into positive and negative sentiment groups. The application of CrossRate is illustrated by its use in estimating the emotions of artworks. It offers even non-art experts the opportunity to form an opinion.

Keywords: convolutional neural network; visual emotion analysis; images of general nature; sentiment analysis; top-2 cross-sentiment measure; art work images; Caravaggio

Check for updates

Academic Editor: Miguel Angel Cazorla

Received: 16 June 2026
Revised: 13 July 2026
Accepted: 31 July 2026
Published: 2 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Nowadays, an increasing amount of data on the Internet is of visual media. Practical value has been demonstrated across various types of images, including medical, satellite, aerial, and remote-sensing imagery. However, a large number of everyday images are also captured snapshots of daily life and travel photos. Recently, it has been reported that more than two in three of the world's population uses social media. It is often difficult to determine what emotion a particular general image conveys. This paper focuses on general-purpose image emotion recognition rather than facial expression recognition. Emotion analysis in general-purpose images began more than 15 years ago.

Early visual emotion recognition (VER) studies relied mainly on handcrafted features, affective concepts, and interpretable visual descriptors. Notable early works include [1,2]. In parallel, several datasets were introduced to support supervised image emotion recognition, including FI-8, ArtPhoto, and EmoSet [3–5]. These early approaches provided interpretable links between visual content and emotion labels, but their recognition performance was limited compared with later deep learning models.

Deep learning-based VER models improved performance by replacing handcrafted representations with learned visual features. Early CNN-based approaches already improved accuracy substantially [6], and subsequent studies developed stronger CNN backbones, attention mechanisms, multiscale region models, and Gram matrix modules [7–13]. More recent methods have also used Transformer-based and multimodal architectures. For example, multiple-view prompting combines image content, generated captions, and enriched emotion labels [14], while dual-branch architectures combine global visual features with visual relationship or region-level features [15]. Other works consider sentiment-level or multimodal emotion analysis [16–19]. These studies show that model architectures for VER have become increasingly powerful, but their evaluation still relies mainly on standard classification performance. Furthermore, recent reviews show that emotion recognition, affective computing, and sentiment analysis are closely related research areas, but they differ in data modalities, applications, and evaluation settings [20,21].

Visual emotion recognition in general-purpose images has wide applications. We review a few below. A unique framework for analyzing emotions at tourist sites using user-generated photos from around the world, combining multi-view emotions and attraction types for a thorough analysis, is presented in [22]. Recognizing and classifying pet emotional states from images is discussed in [23]. Our paper presents another application of VER. More earlier applications are reviewed in [20]. In our paper, the application of VER is also presented. We estimate the emotions of artworks. It is shown that general characteristics, such as colors, textures, and forms, can convey both positive and negative emotions in an artwork. CNN makes it possible for even non-art experts to form an opinion.

Visual emotion recognition is the task of assigning an emotion category to an image, such as amusement, anger, fear, or sadness. However, the emotion associated with an image may vary across annotators [21]. For this reason, many VER datasets collect multiple annotations per image and use majority voting to establish a ground-truth label [3,5]. This enables supervised training, but it may hide disagreement and reduce ambiguous cases to a single category.

Standard supervised classification metrics, such as accuracy, precision, recall, and macro-F1, are used when ground-truth labels are available. These metrics measure agreement with the assigned dataset label, but they do not describe the structure of the model's competing predictions. This is limiting in VER because labels may be subjective, noisy, weakly assigned, or unavailable in practical settings. In particular, these metrics do not indicate whether the two highest-probability emotion classes belong to the same broad affective group or to opposing positive and negative groups.

Label-free uncertainty and confidence measures, such as entropy, maximum softmax probability, and the top-1/top-2 probability margin, can describe how concentrated or uncertain a model's prediction distribution is. The top-1 prediction is the emotion class with the highest predicted probability, while the top-2 prediction is the class with the second-highest predicted probability. However, these measures do not distinguish between uncertainty among emotions within the same sentiment group and uncertainty between opposing sentiment groups. For example, uncertainty between amusement and excitement differs from uncertainty between amusement and fear, even if both cases have similar entropy or top-1/top-2 margins.

These three measures do not distinguish between uncertainty among emotions within the same sentiment group and uncertainty between opposing sentiment groups. For example, uncertainty between amusement and excitement differs from that between amusement and fear, even if both cases have similar entropy or top-1/top-2 margins.

In addition to general confidence and uncertainty measures, ambiguity in visual emotion analysis has also been studied using emotion label distributions and annotator agreement. Visual emotion distribution learning preserves the distribution of annotations across emotion categories rather than reducing multiple responses into a single majority label [24,25]. The similarity between predicted and human-annotated distributions may be evaluated using divergence or distance measures, such as Kullback–Leibler divergence, Jensen–Shannon divergence, or related label-distribution metrics. Annotator disagreement can also be quantified directly using vote entropy or majority agreement. These approaches provide useful measures of human annotation ambiguity, but they require access to individual annotation votes or soft emotion-label distributions.

To address this limitation, our paper proposes the top-2 Cross-Sentiment Rate (CrossRate), a label-free measure of sentiment-level consistency in VER models. CrossRate is a relative measure. When analyzing a given set of images, it indicates the proportion of cases in which the top-1 and top-2 results differ in sentiment. The measure evaluates whether the two highest probability emotion predictions belong to opposite sentiment groups. A lower CrossRate indicates that the model's top-2 predictions tend to remain sentiment-consistent, whereas a higher CrossRate indicates more frequent positive–negative conflicts among them. The proposed measure is evaluated on labeled VER datasets using multiple trained model configurations. In particular, we use contrastive-center loss weighting as an experimental factor to generate model variants with different levels of representation separation. This allows us to evaluate whether the CrossRate measure captures changes in prediction consistency that are not fully shown by standard supervised metrics.

The discussion above identifies three limitations in current VER evaluations. First, standard supervised metrics such as accuracy, precision, recall, and macro-F1 evaluate agreement with ground-truth labels but do not show whether the strongest competing predictions remain within the same sentiment group or cross the positive–negative sentiment boundary. Second, common uncertainty and confidence measures, including entropy, maximum softmax probability, and top-1/top-2 margin, describe the shape of the predictive distribution but do not explicitly account for the sentiment relation between the two strongest emotion predictions. Third, many practical image collections, including artwork datasets, lack reliable ground-truth emotion labels, thereby limiting the use of supervised evaluation metrics.

The main focus of this paper is as follows: to introduce the top-2 cross-sentiment rate for VER, to validate the measure on the labeled VER datasets by comparing its behavior with standard accuracy-based metrics, and to analyze how the proposed measure changes across different training setups. The results demonstrate its practical relevance for analyzing sentiment-level uncertainty in visual emotion predictions. For a given image, the model's top-1 and top-2 predictions can be evaluated at the sentiment level as one of four categories: positive–positive, negative–negative, positive–negative, or negative–positive. Here, e.g., the positive–negative category means that top-1 belongs to the positive sentiment group and top-2 belongs to the negative one. In the case of positive–negative, positive sentiment dominates; in the case of negative–positive, negative sentiment dominates.

The remainder of this paper is organized as follows. Section 2 describes the model architecture, training procedure, and the proposed top-2 cross-sentiment measure (CrossRate). Section 3 presents the experimental results and analysis. Finally, Section 4 concludes the paper and discusses potential directions for future research.

https://doi.org/10.3390/ai7080293

2. Materials and Methods

In this section, we describe the model's structure and the proposed top-2 cross-sentiment measure. Let us note that the proposed measure is appropriate for any model that predicts emotions, which may be grouped into positive and negative sentiments. In this paper, the model is based on a CNN developed in [12,13,26].

2.1. Model Architecture

We use a CNN to extract additional features that improve VER performance. The feature-extraction and Gram matrix module formulation follows our earlier work [26]. To ensure clarity and completeness, we will briefly present the model's architecture in this section, which is described and analyzed in detail in [26]. Let us note that the ideas regarding the CrossRate measure presented in this paper are independent of the chosen model.

Let the convolutional backbone contain L layers, and let $\{l_1, \ldots, l_v\}$, with $v < L$, denote the selected layers from which intermediate feature maps are extracted. For each selected layer l_i, the corresponding feature map is

$$F_{l_i} \in \mathbb{R}^{C_i \times H_i \times W_i}, \tag{1}$$

where C_i, H_i, and W_i denote its number of channels, height, and width, respectively.

Each feature map is processed by a trainable Gram matrix module $g_i(\cdot)$:

$$z_i = g_i(F_{l_i}), \qquad z_i \in \mathbb{R}^{d_i}, \tag{2}$$

with

$$d_i = \lfloor C_i/2 \rfloor. \tag{3}$$

The outputs of the v Gram matrix modules are concatenated into

$$z = [z_1; z_2; \ldots; z_v] \in \mathbb{R}^D, \qquad D = \sum_{i=1}^{v} d_i. \tag{4}$$

Let $h \in \mathbb{R}^{C_*}$ denote the global feature vector obtained from the backbone after global average pooling. To match the dimensionality of the concatenated Gram representation, h is mapped through a learnable projection

$$P : \mathbb{R}^{C_*} \to \mathbb{R}^D, \tag{5}$$

yielding

$$p = P(h), \qquad p \in \mathbb{R}^D. \tag{6}$$

The two feature vectors are combined by element-wise addition

$$u = p + z, \qquad u \in \mathbb{R}^D, \tag{7}$$

where u is the resulting penultimate feature vector. The final prediction is obtained by passing u through the classifier.

In the experiments, feature maps were extracted from three backbone layers with channel dimensions $C_1 = 48$, $C_2 = 64$, and $C_3 = 160$. According to the reduction rule $d_i = \lfloor C_i/2 \rfloor$, the resulting embedding dimensions were $(d_1, d_2, d_3) = (24, 32, 80)$. Therefore, for $v = 3$, the concatenated Gram matrix module representation had dimension $D = 24 + 32 + 80 = 136$. Finally, the feature vectors z, p, and u also have the dimension of $D = 136$.

Figure 1 illustrates the overall architecture of the deep learning model. EfficientNetV2 is employed as the backbone network. It achieved strong classification performance on the ImageNet benchmark [27]. In addition to the final pooled feature vector, the backbone also provides intermediate feature maps from selected layers, which are forwarded to auxiliary Gram matrix modules [12,13].

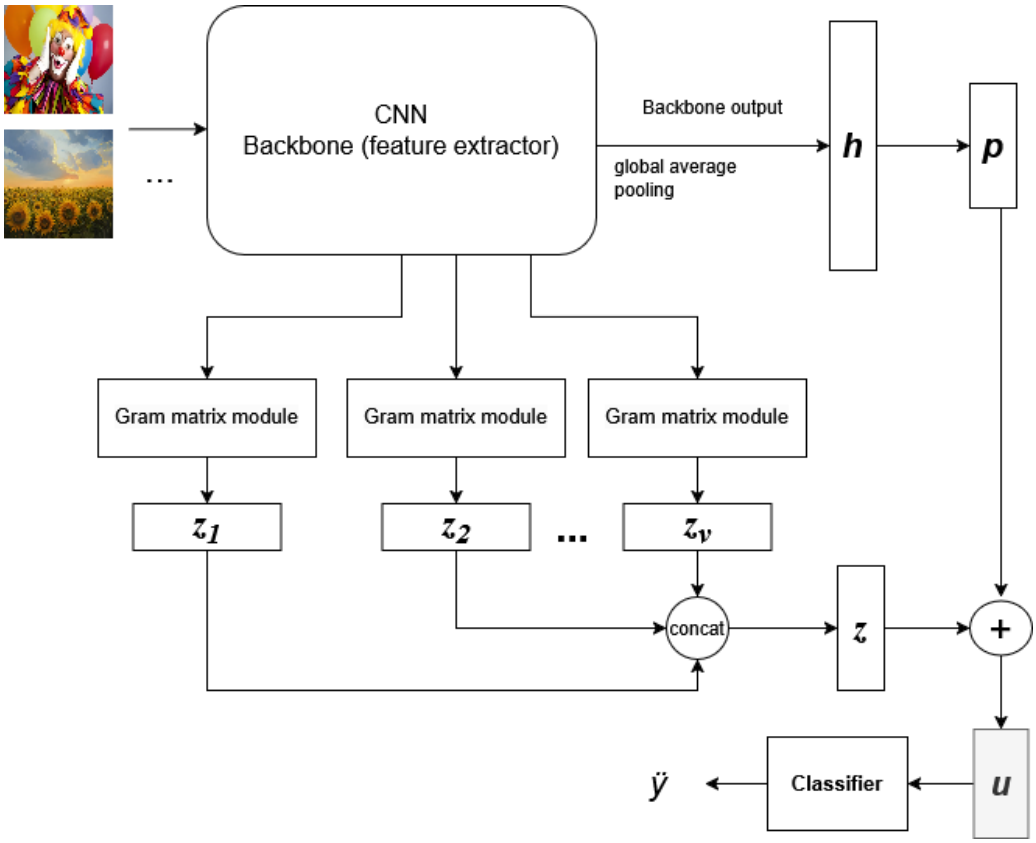

Figure 1. Overall scheme of the model [26].

The Gram matrix $G \in \mathbb{R}^{C \times C}$ is defined in [28] as

$$G = FF^T, \quad F \in \mathbb{R}^{C \times HW}, \quad F^T \in \mathbb{R}^{HW \times C}, \tag{8}$$

where F^T denotes the transpose of F. The matrix F is obtained by reshaping the feature map into a two-dimensional matrix with C rows and HW columns. Therefore, the product FF^T yields a $C \times C$ matrix that captures pairwise correlations between feature channels.

The auxiliary Gram matrix module uses this representation to capture additional structural information from feature activations, thereby improving overall classification performance.

In Figure 2, we can observe the Gram matrix features representation examples generated by the neural transfer style idea [28]. It can be said that Gram matrix features tend to capture low-level visual features with some hints of the structural outline. This also provided visual representations of style, allowing the network to better capture and recognize emotions in images.

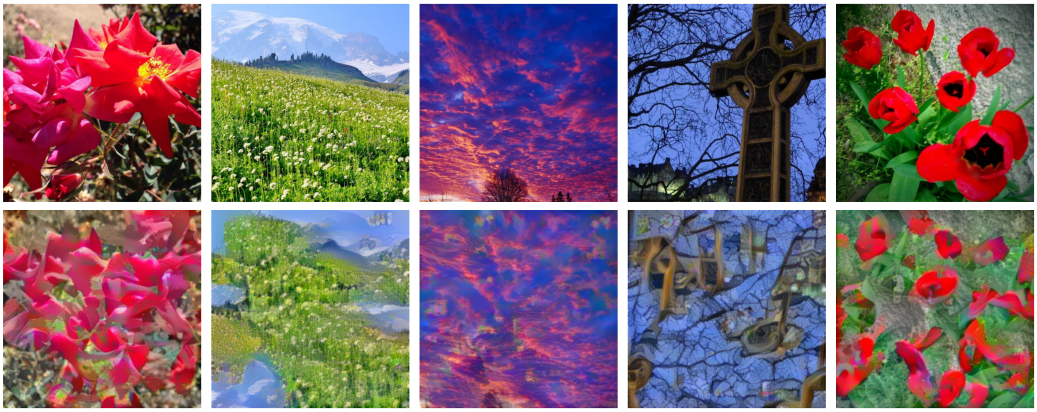

Figure 2. Example of Gram matrix style learned representations.

2.2. Proposed Top-2 Cross-Sentiment Measure–CrossRate

In addition to standard classification metrics, we propose the top-2 cross-sentiment rate (CrossRate), a label-free measure of sentiment-level consistency in the model's prediction distribution. CrossRate instead analyzes whether the two most probable emotion predictions belong to different sentiment groups. A lower CrossRate indicates that the model's strongest competing predictions tend to remain sentiment-consistent, whereas a higher CrossRate indicates more frequent positive–negative conflicts among the top-2 predictions. For each input image x_i, let (c_{1i}, p_{1i}) and (c_{2i}, p_{2i}) denote the top-1 and top-2 predicted emotion classes and their corresponding predicted probabilities, where $p_{1i} \geq p_{2i}$. Let $\mathcal{C}$ be the set of emotion classes. A sentiment mapping is defined as follows:

$$f : \mathcal{C} \to \{\text{positive}, \text{negative}\}.$$

In this study, the positive emotion group contains amusement, awe, contentment, and excitement, while the negative emotion group contains anger, disgust, fear, and sadness. The mapping f is applied according to the class ordering of each dataset.

Depending on the top-2 predictions, the sample is noted as cross-sentiment if

$$f(c_{1i}) \neq f(c_{2i}).$$

For a fixed threshold $\theta \in [0, 0.5)$, only such samples are considered for which the top-2 predicted probability satisfies $p_{2i} \geq \theta$. The eligible set is therefore

$$\mathcal{E}_\theta = \{i : p_{1i} \geq \theta \text{ and } p_{2i} \geq \theta\}, N_\theta = |\mathcal{E}_\theta|.$$

Then the proposed measure is expressed as

$$\text{CrossRate}(\theta) = \frac{1}{N_\theta} \sum_{i \in \mathcal{E}_\theta} \mathbf{1}[f(c_{1i}) \neq f(c_{2i})], \tag{9}$$

where $\mathbf{1}[\cdot]$ is the indicator function. When $\theta = 0$, all samples are eligible, corresponding to the no-threshold version of the measure. For any fixed value of θ with $N_\theta > 0$, CrossRate(θ) lies in the interval $[0, 1]$. A value of 0 means that no eligible samples have cross-sentiment top-2 predictions, whereas a value of 1 means that all eligible samples are cross-sentiment cases. The theoretical range of the measure is independent of the selected threshold.

The confidence threshold is used to focus the analysis on samples with notable probability for both of its two strongest predictions. In particular, thresholding excludes cases where the second class has a very low probability and does not represent a meaningful

alternative. As a result, CrossRate(θ) can capture high-confidence positive–negative or negative–positive ambiguities among the model's top-2 predictions.

CrossRate should be interpreted as a complementary measure rather than a replacement for supervised metrics. When analyzing a given set of images, it indicates the proportion of cases in which the top-1 and top-2 results differ in sentiment. A low CrossRate does not guarantee that the fine-grained emotion prediction is correct; it only indicates that the model's top-2 predictions are sentiment-consistent. Conversely, a high CrossRate may reflect model uncertainty, image ambiguity, noisy labels, or conflicting visual features.

The threshold θ controls how strong the second-ranked prediction must be in order for a sample to be included in the thresholded CrossRate calculation. Since $p_{1i} \geq p_{2i}$, requiring both $p_{1i} \geq \theta$ and $p_{2i} \geq \theta$ mainly filters samples according to the probability of the second-ranked class. Therefore, the thresholded version of CrossRate focuses on cases where the second prediction represents a meaningful competing emotion rather than a low-probability residual class.

In this study, $\theta = 0.3$ was selected as the main thresholded setting because it provides a practical balance between inclusiveness and strictness. Lower thresholds, such as $\theta = 0.1$, include many samples where the second prediction (top-2) may still be relatively weak, while higher thresholds, such as $\theta = 0.4$ or $\theta = 0.5$, are too restrictive for an eight-class softmax output and may exclude many relevant top-2 ambiguity cases. A threshold of $\theta = 0.3$ ensures that the top-2 emotion prediction receives a substantial probability mass, while still keeping enough eligible samples for meaningful analysis. The value $\theta = 0.2$ is also a reasonable alternative for a less restrictive analysis, but $\theta = 0.3$ was used as the primary threshold option.

The proposed top-2 cross-sentiment measure is useful for individual image instances. For a given image, the model's top-1 and top-2 emotion predictions can be assigned at the sentiment level as one of four cases: positive–positive, negative–negative, positive–negative, or negative–positive. The first two outcomes indicate sentiment-consistent predictions, while the last two indicate cross-sentiment predictions. In the cross-sentiment cases, the dominant sentiment is defined by the top-1 emotion prediction; therefore, a positive–negative case can be viewed as predominantly positive, and a negative–positive case as predominantly negative.

2.3. Uncertainty and Confidence Indicators

In addition to the proposed CrossRate measure, we also compute several standard label-free uncertainty and confidence indicators from the model's predicted probability distribution. These indicators are used as comparison measures to analyze whether CrossRate captures information distinct from general predictive uncertainty.

For each input image x_i, let the model outputs a set of probabilities $p_i = (p_{i1}, p_{i2}, \ldots, p_{iE})$ over E emotions, where $E = 8$ in this study. Let the members of p_i be ordered in decreasing order. Then p_{i1} and p_{i2} are the highest and second-highest predicted probabilities, respectively. The following measures are computed for each sample:

$$\begin{aligned} H_i &= -\sum_{j=1}^{E} p_{ij} \log(p_{ij}), \\ H_i^{norm} &= \frac{H_i}{\log(E)}, \\ MSP_i &= p_{i1}, \\ Margin_i &= p_{i1} - p_{i2}. \end{aligned} \tag{10}$$

Here, H_i is the predictive entropy. It measures how widely the predicted probability mass is distributed across emotion classes. A low entropy value indicates that the model's prediction is concentrated on one or a few classes, while a high entropy value indicates that the model distributes probability more evenly across several classes [29]. The normalized entropy H_i^{norm} rescales entropy by its maximum possible value, $\log(E)$, making it easier to compare values across settings [29].

The maximum softmax probability, MSP_i, is the probability assigned to the top-1 predicted emotion class [30]. A higher MSP value indicates stronger confidence in the model's selected class. The margin measures the difference between the top-1 and top-2 predicted probabilities. A large margin indicates that the model prefers its top prediction, whereas a small margin indicates that the top two emotion classes are competing closely.

For a test set of N samples, we report the mean values of the following indicators: mean entropy, mean normalized entropy, mean MSP, and mean margin. These aggregated values summarize the overall uncertainty and confidence of a model on a given dataset [29].

These measures are relevant to VER because an image may carry more than one emotional response. For example, the same image may contain visual features that are related to both contentment and sadness, or to both awe and fear. In such cases, entropy and margin help describe whether the model strongly prefers one emotion class or whether probability is distributed across several competing classes.

However, these uncertainty indicators do not use the semantic structure of the emotion classes. They describe the overall confidence of the prediction distribution, but they do not distinguish between competition among emotions of the same sentiment group and competition among emotions of opposite sentiment groups. For example, a model may be uncertain whether to choose amusement or excitement, or amusement or fear. These two cases may have similar entropy, MSP, or margin values, but they have different meanings for visual emotion recognition. The first case remains within the positive group, where the second crosses the positive–negative boundary. Therefore, entropy, MSP, and margin are used as general uncertainty baselines, while CrossRate specifically evaluates whether the model's top-2 predictions are sentiment-consistent.

2.4. Datasets

The dataset selection and preprocessing procedure followed our earlier work [26]. For completeness, the main characteristics of the datasets used in the present study are summarized below.

The Flickr and Instagram dataset (FI-8) [3] was used for model training and evaluation. It contains 23,308 images of the eight emotion categories [31]. The images were collected from Flickr and Instagram using emotion-related queries. Each image was then evaluated by five Amazon Mechanical Turk annotators. Only images that received more than three votes for the same emotion category were retained.

The second dataset for visual emotion analysis is EmoSet-118K [5]. It contains 118,102 manually annotated images.

The class distributions of FI-8 and EmoSet-118K are reported in Table 1.

EmoSet-118K has a well-balanced distribution across the four positive categories: amusement, awe, contentment, and excitement, and the four negative categories: anger, disgust, fear, and sadness. Each image in EmoSet-118K was assessed by 10 human annotators, and a final label was retained when more than seven annotators selected the same emotion category. Amusement and excitement are the most represented classes, whereas anger and disgust contain the fewest images. In FI-8, the smallest classes are anger, disgust, and fear.

Table 1. Content of FI-8 and EmoSet-118K datasets [26].

Positive	Amusement	Awe	Contentment	Excitement
EmoSet-118K	19,445	15,037	16,337	19,828
FI-8	4924	3151	5374	2963
Negative	Anger	Disgust	Fear	Sadness
EmoSet-118K	10,660	10,666	13,453	12,676
FI-8	1266	1658	1032	2922

In addition, the WikiArt emotions dataset was used for evaluation [32]. This dataset consists of artwork images annotated with the emotions they evoke in viewers. The original WikiArt labels were not used. Instead, the trained model was applied to 3176 WikiArt images, producing an eight-class emotion prediction for each image. The proposed CrossRate measure was then computed from the model's top-1 and top-2 predicted emotion classes after mapping these predictions to positive and negative sentiment groups. Therefore, WikiArt was used to assess the applicability of CrossRate in a setting where fine-grained ground-truth labels for the model's eight emotion classes are unavailable.

The training, validation, and test splits provided by the authors of EmoSet-118K were kept unchanged. The FI-8 dataset was split into 85% training and 15% test subsets.

2.5. Model Training Loss

The loss formulation used for model training follows our earlier work [26]. For clarity and completeness, its main components are summarized below. The model is trained using sparse categorical cross-entropy loss and contrastive-center loss. Sparse categorical cross-entropy is used for the main classification objective, and contrastive-center loss is applied to the penultimate feature representation.

Let E be the number of visual emotion classes, and let $\mathbf{y} = (y_1, \ldots, y_N)$ denote the ground-truth label vector for N samples, with $y_k \in \{1, \ldots, E\}$. The sparse categorical cross-entropy loss is defined as

$$L_{\text{entr}} = -\frac{1}{N} \sum_{k=1}^{N} \log(p_{k,y_k}), \tag{11}$$

where p_{k,y_k} denotes the model-predicted probability of the ground-truth class y_k for sample k.

2.6. Contrastive-Center Loss

Contrastive-center loss is computed from the penultimate feature vectors of the model. Let $\mathbf{x}_k \in \mathbb{R}^{136}$ denote the feature vector of k-th sample, and let $\mathbf{c}_1, \ldots, \mathbf{c}_E$ denote the learnable class centers. Each class center has the same dimensionality as $\mathbf{x}_k$ and is updated during training together with the model parameters.

The contrastive-center loss consists of two components:

$$L_{\text{contr}} = L_{\text{intra}} + \lambda L_{\text{inter}}, \tag{12}$$

where L_{intra} enforces samples of the same class to be close to their class center, and L_{inter} enforces different class centers to be separated. The parameter λ controls the contribution of the inter-class separability term.

The intra-class compactness component is defined as

$$L_{\text{intra}} = \frac{1}{N} \sum_{k=1}^{N} \|\mathbf{x}_k - \mathbf{c}_{y_k}\|^2, \tag{13}$$

where $\mathbf{c}_{y_k}$ is the class center corresponding to the ground-truth label y_k. This term minimizes the squared Euclidean distance between each feature vector and its corresponding class center.

The inter-class separability component is defined as

$$L_{\text{inter}} = \frac{1}{E(E-1)} \sum_{i=1}^{E} \sum_{j \neq i} \max(0, m - \|\mathbf{c}_i - \mathbf{c}_j\|)^2, \tag{14}$$

where m is the margin that enforces a minimum distance between different class centers. If the distance between two centers is greater than or equal to m, no penalty is applied. If the distance is smaller than m, the loss penalizes this pair of centers.

Integration with Cross-Entropy Loss

The total training loss combines sparse categorical cross-entropy and contrastive-center loss:

$$L_{\text{total}} = L_{\text{entr}} + \beta L_{\text{contr}}, \tag{15}$$

where L_{entr} is computed from the classifier output, and L_{contr} is computed from the penultimate feature vectors.

The parameter β controls the contribution of the contrastive-center loss. When $\beta = 0$, the model is trained using only cross-entropy loss and serves as the baseline. Larger values of β place greater emphasis on feature compactness and class-center separation.

2.7. Experimental Setup

The training procedure was adapted from our earlier work [26]. The main hyperparameters and augmentation settings used in the present study are summarized below.

Two model configurations were trained, one for each visual emotion dataset. The dataset-specific hyperparameters were the learning rate, number of training epochs, β, margin m, and δ. Training was performed using stochastic gradient descent (SGD) with a momentum of 0.9 and an initial learning rate of 0.02. The learning rate was adjusted using cosine annealing with warm restarts [33], decaying from 0.02 to 1×10^{-4} before restarting every five epochs. The parameter β controlled the contribution of the contrastive-center loss. The margin m was set to 5, and λ was set to 1. During training, each image was first subjected to a random resized crop to 224×224 and a random horizontal flip. RandAugment [34] was then applied with a distortion strength of three and two transformations. Following the procedure used in [26], gradient clipping was applied to control the gradient instability caused by the contrastive-center loss.

The model was implemented and trained using the PyTorch 2.6.0 deep learning library. Training was performed for 20 epochs using a distributed setup with 2× V100 GPUs. The total batch size was 256, split evenly across the two GPUs.

3. Results

We use the CrossRate measure, as described in Equation (9), to evaluate the model's internal consistency.

Initial top-2 cross-sentiment results were reported in our earlier work [26]. Table 2 presents the evaluation on the WikiArt emotions dataset across several confidence thresholds. Eligible denotes the number of samples for which both top-1 and top-2 probabilities satisfy the selected threshold. The proposed approach, which uses contrastive-center loss, yields a lower CrossRate than the baseline at every threshold. As θ increases, fewer samples remain eligible, because cases with a weak top-2 prediction are excluded.

Table 2. Top-2 cross-sentiment evaluation results. CrossRate is given in percent.

θ	Model	Eligible	pos-neg	neg-pos	pos-pos	neg-neg	CrossRate
No threshold	Baseline	3176	652	664	518	1342	41.4
	Proposed approach	3176	288	326	894	1668	19.3
0.10	Baseline	2024	443	439	317	825	43.6
	Proposed approach	1412	200	247	306	659	31.7
0.20	Baseline	1260	278	279	196	507	44.2
	Proposed approach	753	99	133	163	358	30.8
0.30	Baseline	552	122	121	85	231	43.5
	Proposed approach	277	22	33	59	163	19.9
0.40	Baseline	110	21	23	15	51	40.0
	Proposed approach	41	2	2	6	31	9.8

Table 3 presents the supervised classification performance of the evaluated models on the EmoSet-118K test set. The best fine-grained emotion recognition performance is obtained with $\beta = 0.5$, which achieves the highest accuracy, macro-F1, macro-precision, and macro-recall. Increasing β beyond 0.5 does not improve standard supervised metrics, suggesting that stronger contrastive-center weighting does not necessarily lead to better eight-class classification performance.

Table 3. Classification performance on the EmoSet-118K test set for different values of the contrastive-center weighting coefficient β.

β	Accuracy (%)	Macro-F1	Macro-Precision	Macro-Recall	Sentiment Accuracy (%)
0.0	79.19	0.7989	0.8017	0.7966	95.08
0.1	80.18	0.8092	0.8132	0.8066	95.69
0.5	**80.42**	**0.8117**	**0.8161**	**0.8089**	95.68
0.7	80.24	0.8090	0.8122	0.8069	**95.76**
1.0	80.08	0.8078	0.8121	0.8048	95.65

The best result in each metric column is shown in bold.

Table 4 shows the effect of β on top-2 cross-sentiment consistency. In contrast to the standard supervised metrics, CrossRate decreases substantially as β increases. The baseline model with $\beta = 0.0$ produces cross-sentiment top-2 predictions for 22.62% of test samples, while the model with $\beta = 1.0$ reduces this value to 6.08%. This indicates that stronger contrastive-center weighting substantially reduces positive–negative conflicts among the model's two most probable predictions. Under the confidence threshold $\theta = 0.3$, the lowest CrossRate is obtained at $\beta = 0.7$, although the difference between $\beta = 0.7$ and $\beta = 1.0$ is very small.

Table 4. Top-2 cross-sentiment results on the EmoSet-118K test set for different values of β. CrossRate is reported both without a confidence threshold and with $\theta = 0.3$.

β	CrossRate ($\theta = 0$)	Cross Cases ($\theta = 0$)	CrossRate ($\theta = 0.3$)	Cross Cases ($\theta = 0.3$)	Eligible ($\theta = 0.3$)
0.0	22.62%	4007	15.84%	369	2330
0.1	14.63%	2592	12.44%	235	1889
0.5	7.04%	1248	9.32%	132	1417
0.7	6.46%	1144	**8.32%**	110	1322
1.0	**6.08%**	1077	8.34%	**102**	1223

The best result in each metric column is shown in bold.

Table 5 reports general uncertainty and confidence indicators. These values change only moderately across different β settings. Mean entropy and normalized entropy increase slightly as β increases, while MSP and margin remain relatively stable and high. Therefore, the large reduction in CrossRate observed in Table 4 cannot be explained simply by a major change in overall model confidence. Instead, the results suggest that β mainly affects the sentiment-level organization of the top-2 predictions rather than merely increasing or decreasing the model's general confidence.

Table 5. Mean uncertainty and confidence indicators on the EmoSet-118K test set for different values of β. Entropy and normalized entropy measure the dispersion of the predicted probability distribution, MSP denotes the maximum softmax probability, and margin denotes the difference between the top-1 and top-2 predicted probabilities.

β	Mean Entropy	Mean Normalized Entropy	Mean MSP	Mean Margin
0.0	0.407	0.196	0.848	0.738
0.1	0.431	0.207	0.851	0.753
0.5	0.462	0.222	0.860	0.776
0.7	0.465	0.224	**0.861**	**0.779**
1.0	**0.477**	**0.229**	0.859	0.778

The best result in each metric column is shown in bold.

Tables 6–8 present repeated-seed results for the baseline model ($\beta = 0.0$) and the selected contrastive-center loss setting ($\beta = 0.5$). Each configuration was trained five times using different fixed random seeds. The results show that $\beta = 0.5$ consistently improves supervised performance compared with the baseline. Accuracy increases from $79.13 \pm 0.16\%$ to $80.10 \pm 0.15\%$, while Macro-F1 increases from 0.7982 ± 0.0007 to 0.8083 ± 0.0016. At the same time, CrossRate decreases substantially, from $22.15 \pm 0.45\%$ to $7.81 \pm 0.62\%$ without thresholding and from $16.83 \pm 0.74\%$ to $8.66 \pm 1.07\%$ with $\theta = 0.3$. The uncertainty and confidence indicators also change, but less notably: entropy and margin increase under $\beta = 0.5$, while MSP remains high in both settings. These results indicate that the CrossRate reduction is stable across training runs and is not solely attributable to a simple decrease in overall model confidence.

Table 6. Repeated-seed classification performance on the EmoSet-118K test set for the baseline model ($\beta = 0.0$) and the selected contrastive-center loss setting ($\beta = 0.5$). Results are reported as mean $\pm$ standard deviation over five training runs with different fixed random seeds.

β	Accuracy (%)	Macro-F1	Macro-Precision	Macro-Recall	Sentiment Accuracy
0.0	79.13 ± 0.16	0.7982 ± 0.0007	0.8013 ± 0.0009	0.7958 ± 0.0010	0.9528 ± 0.0013
0.5	**80.10** ± 0.15	**0.8083** ± 0.0016	**0.8117** ± 0.0020	**0.8059** ± 0.0016	**0.9567** ± 0.0009

The best result in each metric column is shown in bold.

Table 7. Repeated top-2 cross-sentiment results on the EmoSet-118K test set for the baseline model ($\beta = 0.0$) and the selected contrastive-center loss setting ($\beta = 0.5$). Results are reported as mean $\pm$ standard deviation over five training runs with different fixed random seeds.

β	CrossRate ($\theta = 0$)	Cross Cases ($\theta = 0$)	CrossRate ($\theta = 0.3$)	Cross Cases ($\theta = 0.3$)	Eligible ($\theta = 0.3$)
0.0	$22.15 \pm 0.45\%$	3924 ± 80	$16.83 \pm 0.74\%$	374 ± 15	2222 ± 46
0.5	**7.81** $\pm 0.62\%$	**1384** ± 109	**8.66** $\pm 1.07\%$	**124** ± 17	**1428** ± 31

The best result in each metric column is shown in bold.

Table 9 presents the classification performance on the FI-8 test set. The best performance is achieved with $\beta = 0.5$, which gives the highest accuracy, macro-F1, macro-precision, and macro-recall. Increasing β beyond 0.5 does not further improve the accuracy-

based results, suggesting that stronger contrastive-center weighting may already be reducing fine-grained emotion discrimination on FI-8.

Table 8. Repeated-seed mean uncertainty and confidence indicators on the EmoSet-118K test set for the baseline model ($\beta = 0.0$) and the selected contrastive-center loss setting ($\beta = 0.5$). Results are reported as mean ± standard deviation over five training runs with different fixed random seeds.

β	Mean Entropy	Mean Normalized Entropy	Mean MSP	Mean Margin
0.0	0.3919 ± 0.0027	0.1885 ± 0.0013	0.8521 ± 0.0011	0.7448 ± 0.0016
0.5	**0.4693** ± 0.0065	**0.2257** ± 0.0031	**0.8567** ± 0.0019	**0.7720** ± 0.0030

The best result in each metric column is shown in bold.

Table 9. Supervised classification performance on the FI-8 test set for different values of the contrastive-center weighting coefficient β.

β	Accuracy (%)	Macro-F1	Macro-Precision	Macro-Recall	Sentiment Accuracy (%)
0.0	69.80	0.6421	0.6579	0.6303	89.20
0.1	71.82	0.6632	0.6807	0.6507	89.37
0.5	**71.94**	**0.6698**	**0.6827**	**0.6612**	90.26
0.7	71.73	0.6664	0.6799	0.6567	**90.75**
1.0	71.68	0.6564	0.6679	0.6479	90.40

The best result in each metric column is shown in bold.

Table 10 reports the top-2 cross-sentiment results on FI-8. As β increases, the unthresholded CrossRate decreases consistently from 26.30% at $\beta = 0.0$ to 19.46% at $\beta = 1.0$. This indicates that stronger contrastive-center weighting reduces positive–negative conflicts between the model's two most probable predictions. Under the confidence threshold $\theta = 0.3$, the lowest CrossRate is obtained at $\beta = 0.7$, while $\beta = 1.0$ produces the smallest number of high-confidence cross-sentiment cases. Overall, the FI-8 results show that CrossRate improves with larger β, although high-confidence cross-sentiment ambiguity remains more frequent than in the EmoSet-118K experiments.

Table 10. Top-2 cross-sentiment results on the FI-8 test set for different values of β. CrossRate is reported both without a confidence threshold and with $\theta = 0.3$.

β	CrossRate ($\theta = 0$)	Cross Cases ($\theta = 0$)	CrossRate ($\theta = 0.3$)	Cross Cases ($\theta = 0.3$)	Eligible ($\theta = 0.3$)
0.0	26.30%	896	26.72%	132	494
0.1	24.07%	820	26.37%	101	383
0.5	21.19%	722	23.56%	78	331
0.7	19.87%	677	**23.24%**	76	327
1.0	**19.46%**	**663**	23.55%	**73**	310

The best result in each metric column is shown in bold.

Table 11 summarizes general uncertainty and confidence indicators on FI-8. Unlike CrossRate, these indicators do not change monotonically with β. Mean entropy is highest at $\beta = 0.1$, while MSP and margin remain within a relatively narrow range across the evaluated configurations. This suggests that changes in CrossRate are not solely due to a uniform increase or decrease in model confidence. Instead, CrossRate captures changes in the sentiment-level structure of the top-2 predictions that are not fully reflected by entropy, MSP, or margin.

Tables 12–14 present repeated-seed results on the FI-8 test set for the baseline model ($\beta = 0.0$) and the selected contrastive-center loss setting ($\beta = 0.5$). Each configuration was trained five times using different fixed random seeds. The results show that $\beta = 0.5$ improves classification performance compared with the baseline. Accuracy increases

from $69.64 \pm 0.37\%$ to $71.76 \pm 0.27\%$, while Macro-F1 increases from 0.6441 ± 0.0043 to 0.6629 ± 0.0045. CrossRate also decreases, from $27.08 \pm 0.78\%$ to $22.37 \pm 0.80\%$ without thresholding and from $29.15 \pm 1.42\%$ to $23.42 \pm 1.63\%$ with $\theta = 0.3$. In contrast to the EmoSet-118K results, entropy increases while MSP and margin decrease under $\beta = 0.5$, suggesting that the model becomes less confident on FI-8 despite achieving better classification performance and lower CrossRate. These results indicate that the CrossRate reduction is stable across training runs, although the changes between CrossRate and general confidence indicators vary across datasets.

Table 11. Mean uncertainty and confidence indicators on the FI-8 test set for different values of β. Entropy and normalized entropy measure the dispersion of the predicted probability distribution, MSP denotes the maximum softmax probability, and margin denotes the difference between the top-1 and top-2 predicted probabilities.

β	Mean Entropy	Mean Normalized Entropy	Mean MSP	Mean Margin
0.0	0.496	0.238	**0.817**	0.691
0.1	**0.630**	**0.303**	0.788	0.661
0.5	0.592	0.285	0.808	0.696
0.7	0.606	0.292	0.804	0.691
1.0	0.594	0.286	0.811	**0.701**

The best result in each metric column is shown in bold.

Table 12. Repeated-seed classification performance on the FI-8 test set for the baseline model ($\beta = 0.0$) and the selected contrastive-center loss setting ($\beta = 0.5$). Results are reported as mean ± standard deviation over five training runs with different fixed random seeds.

β	Accuracy (%)	Macro-F1	Macro-Precision	Macro-Recall	Sentiment Accuracy
0.0	69.64 ± 0.37	0.6441 ± 0.0043	0.6581 ± 0.0041	0.6338 ± 0.0050	0.8862 ± 0.0026
0.5	**71.76** ± 0.27	**0.6629** ± 0.0045	**0.6852** ± 0.0039	**0.6490** ± 0.0053	**0.9027** ± 0.0020

The best result in each metric column is shown in bold.

Table 13. Repeated-seed top-2 cross-sentiment results on the FI-8 test set for the baseline model ($\beta = 0.0$) and the selected contrastive-center loss setting ($\beta = 0.5$). Results are reported as mean ± standard deviation over five training runs with different fixed random seeds.

β	CrossRate ($\theta = 0$)	Cross Cases ($\theta = 0$)	CrossRate ($\theta = 0.3$)	Cross Cases ($\theta = 0.3$)	Eligible ($\theta = 0.3$)
0.0	$27.08 \pm 0.78\%$	923 ± 27	$29.15 \pm 1.42\%$	139 ± 7	477 ± 13
0.5	**22.37** $\pm 0.80\%$	**762** ± 27	**23.42** $\pm 1.63\%$	**79** ± 8	338 ± 19

The best result in each metric column is shown in bold.

Table 14. Repeated-seed mean uncertainty and confidence indicators on the FI-8 test set for the baseline model ($\beta = 0.0$) and the selected contrastive-center loss setting ($\beta = 0.5$). Results are reported as mean ± standard deviation over five training runs with different fixed random seeds.

β	Mean Entropy	Mean Normalized Entropy	Mean MSP	Mean Margin
0.0	0.5601 ± 0.0096	0.2693 ± 0.0046	**0.7967** ± 0.0040	**0.6638** ± 0.0065
0.5	**0.7470** ± 0.0112	**0.3592** ± 0.0054	0.7593 ± 0.0040	0.6294 ± 0.0062

The best result in each metric column is shown in bold.

3.1. Practical Case Study on the Artwork Images

Figure 3 presents several artwork images to illustrate the proposed sentiment-level analysis. The first artwork is attributed to Michelangelo Merisi da Caravaggio—Mary Magdalene in Ecstasy, c. 1606, the second to Alberto Magnelli—Lyrical Explosion No. 12, 1918, the third to Eyvind Earle—Steep Coast Line, 1996, and the fourth to James Weeks—Looking West from Spanish Fort–Baker Beach #3, 1962.

Emotion recognition in general-purpose images is illustrated in this section. Rather than performing a baseline comparison, the analysis is intended to demonstrate the applicability of the CrossRate to images for which no prior emotion labels are available. The model, trained on EmoSet-118K, was applied to the images in Figure 3. The model's performance was optimized for top-2 cross-sentiment analysis applying $\beta = 0.5$. As a result, our model gives probabilities of all eight emotions for each artwork image. Each image is evaluated separately using the proposed top-2 cross-sentiment measure. Based on the sentiment categories of these two emotions, the final sentiment prediction is assigned as one of four cases: negative sentiment, positive sentiment, negative sentiment dominates, or positive sentiment dominates. Sentiment predictions are given above each picture. Domination means that the two top emotions are from different sentiments. Domination is determined by the top-1 emotion prediction.

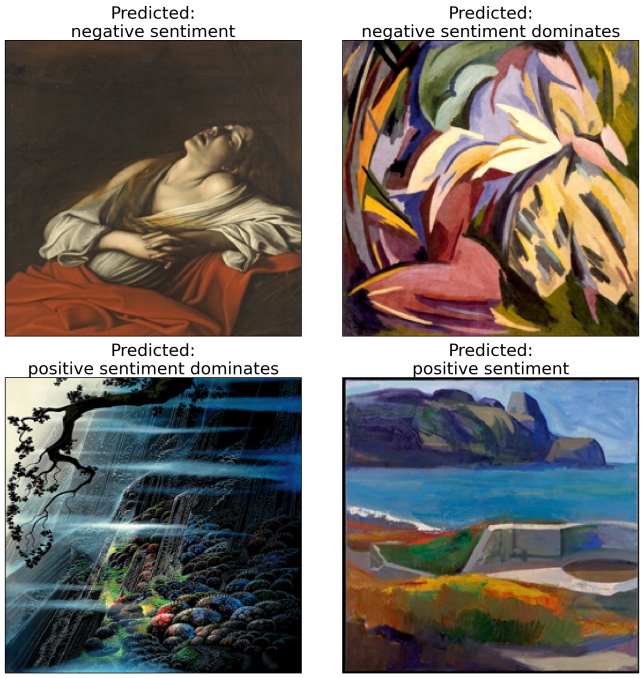

Figure 3. Visual emotion predictions and sentiment-level interpretation for artwork images.

For details, the top-1 and top-2 emotion predictions are presented below. Michelangelo Merisi da Caravaggio: fear (probability 0.330), sadness (0.291). Alberto Magnelli: fear (0.411), amusement (0.219). Eyvind Earle: awe (0.690), fear (0.160). James Week: awe (0.960), contentment (0.023).

In our given example, we classify four artworks using the trained network. Above each image, we present the predicted sentiment or its dominance. The model predicts negative sentiment for the first artwork and positive sentiment for the fourth artwork. For the second artwork, the prediction is less certain, but negative sentiment still dominates. Furthermore, the third artwork is also predicted to be a mixed case with positive sentiment dominance.

3.2. Limitations of the Study

Several limitations should be noted. First, although CrossRate is label-free with respect to ground-truth emotion annotations, it requires that predicted emotion classes be mapped to sentiment groups. This mapping is straightforward for the eight emotion classes used in this study, but it may be less clear for datasets that include neutral, mixed, or ambiguous emotion categories. Second, the experiments are based on single-label emotion annotations and therefore do not evaluate annotation-based ambiguity measures, such as annotator agreement, vote entropy, or sentiment-distribution divergence. Third, although

an additional ablation was included, the study does not provide a full architecture-level comparison across alternative backbones such as transformer-based VER models.

4. Conclusions

This paper proposes the top-2 cross-sentiment rate (CrossRate) as a label-free complementary measure for visual emotion recognition models, aiming to analyze sentiment-level ambiguities. The measure is relevant because standard supervised metrics do not indicate whether the model's strongest competing-emotion predictions remain within the same positive or negative sentiment group or cross the sentiment boundary. The experiments showed that the proposed measure changes across different model configurations and decreases as the contrastive-center loss weight β increases, especially on EmoSet-118K and FI-8. These changes were not fully displayed by commonly used uncertainty measures. The WikiArt case study also demonstrated that the measure can be applied even when ground-truth emotion labels are unavailable.

The results have a clearly recognizable practical significance. When applying the model to the particular image, its sentiment is predicted as one of four cases: negative sentiment, positive sentiment, negative sentiment dominates, or positive sentiment dominates.

Nevertheless, the model used in this paper is based on CNN EfficientNetV2S expanded with the Gram matrix modules [12,13], the proposed measure is appropriate for any model that predicts emotions, which may be grouped into positive and negative sentiments.

General characteristics such as colors, textures, and forms can convey both positive and negative emotions in an artwork. Therefore, the application of CrossRate is reasonable for estimating the emotions of artworks. In this paper, such an application is provided. It offers even non-art experts the opportunity to form an opinion.

As discussed in the limitations of the study subsection, CrossRate depends on the availability of a sentiment mapping and was not compared with annotation-based ambiguity measures such as annotator agreement or sentiment-distribution divergence.

Future work might include studies to evaluate the proposed measure across a diversity of datasets and model architectures.

Author Contributions: Conceptualization, M.M. and G.D.; methodology, M.M. and G.D.; software, M.M.; validation, M.M. and G.D.; formal analysis, G.D.; investigation, M.M. and G.D.; data curation, M.M.; writing—original draft preparation, M.M. and G.D.; writing—review and editing, M.M. and G.D.; visualization, M.M. and G.D.; supervision, G.D.; project administration, G.D.; funding acquisition, G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available for EmoSet at https://vcc.tech/EmoSet (accessed on 30 July 2026), WikiArt Emotions at https://saifmohammad.com/WebPages/wikiartemotions.html (accessed on 30 July 2026). The FI-8 dataset was introduced by You et al. [3]; its availability is subject to the conditions specified by the original dataset providers.

Acknowledgments: The authors are thankful to the Information Technology Research Center of Vilnius University for the high-performance computing resources.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CrossRate	Top-2 Cross-Sentiment Measure
CNN	Convolutional Neural Network
MSP	Maximum Softmax Probability
VEA	Visual Emotion Analysis
VER	Visual Emotion Recognition

References

1. Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013*; ACM: New York, NY, USA, 2013; pp. 223–232. [CrossRef]
2. Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.S.; Sun, X. Exploring principles-of-art features for image emotion recognition. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 47–56.
3. You, Q.; Luo, J.; Jin, H.; Yang, J. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *arXiv* **2016**, arXiv:1605.02677. [CrossRef]
4. Panda, R.; Zhang, J.; Li, H.; Lee, J.Y.; Lu, X.; Roy-Chowdhury, A.K. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the Computer Vision–ECCV, Munich, Germany, 8–14 September 2018*; Vittorio, F., Martial, H., Cristian, S., Yair, W., Eds.; Springer: Cham, Switzerland, 2018; pp. 594–612.
5. Yang, J.; Huang, Q.; Ding, T.; Lischinski, D.; Cohen-Or, D.; Huang, H. EmoSet: A large-scale visual emotion dataset with rich attributes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 20326–20337. [CrossRef]
6. Chen, T.; Borth, D.; Darrell, T.; Chang, S.F. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv* **2014**, arXiv:1410.8586.
7. Riaz, M.N.; Shen, Y.; Sohail, M.; Guo, M. eXnet: An efficient approach for emotion recognition in the wild. *Sensors* **2020**, *20*, 1087. [CrossRef] [PubMed]
8. Yang, J.; She, D.; Sun, M. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; pp. 3266–3272.
9. Tan, L.; Zhang, K.; Wang, K.; Zeng, X.; Peng, X.; Qiao, Y. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In *Proceedings of the ICMI 2017–Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017*; Association for Computing Machinery, Inc.: New York, NY, USA, 2017; pp. 549–552. [CrossRef]
10. Karbauskaite, R.; Sakalauskas, L.; Dzemyda, G. Kriging predictor for facial emotion recognition using numerical proximities of human emotions. *Informatica* **2020**, *31*, 249–275. [CrossRef]
11. Zhang, H.; Xu, D.; Luo, G.; He, K. Learning multi-level representations for affective image recognition. *Neural Comput. Appl.* **2022**, *34*, 14107–14120. [CrossRef]
12. Motiejauskas, M.; Dzemyda, G. The effective evaluation of emotions in the visual emotion images using convolutional neural networks. *IEEE Access* **2025**, *13*, 139174–139187. [CrossRef]
13. Motiejauskas, M.; Dzemyda, G. EfficientNet convolutional neural network with Gram matrices modules for predicting sadness emotion. *Int. J. Comput. Commun. Control* **2024**, *19*, 6697. [CrossRef]
14. Xu, Q.; Wei, Y.; Yuan, S.; Wu, J.; Wang, L.; Wu, C. Learning emotional prompt features with multiple views for visual emotion analysis. *Inf. Fusion* **2024**, *108*, 102366. [CrossRef]
15. Luo, Y.; Zhong, X.; Xie, J.; Liu, G. CVRSF-Net: Image emotion recognition by combining visual relationship features and scene features. *IEEE Trans. Emerg. Top. Comput. Intell.* **2025**, *9*, 2321–2333. [CrossRef]
16. Zhang, H.; Xu, M. Multiscale Emotion Representation Learning for Affective Image Recognition. *IEEE Trans. Multimed.* **2023**, *25*, 2203–2212. [CrossRef]
17. Xu, Q.; Yuan, S.; Wei, Y.; Wu, J.; Wang, L.; Wu, C. Multiple feature refining network for visual emotion distribution learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 8924–8932.
18. Xu, L.; Wang, Z.; Wu, B.; Lui, S. MDAN: Multi-level Dependent Attention Network for Visual Emotion Analysis. *arXiv* **2022**, arXiv:2203.13443.
19. Rui, X. A convolutional neural networks based approach for clustering of emotional elements in art design. *PeerJ Comput. Sci.* **2023**, *9*, e1548. [CrossRef] [PubMed]

20. García-Hernández, R.A.; Luna-García, H.; Celaya-Padilla, J.M.; García-Hernández, A.; Reveles-Gómez, L.C.; Flores-Chaires, L.A.; Delgado-Contreras, J.R.; Rondon, D.; Villalba-Condori, K.O. A Systematic Literature Review of Modalities, Trends, and Limitations in Emotion Recognition, Affective Computing, and Sentiment Analysis. *Appl. Sci.* **2024**, *14*, 7165. [CrossRef]
21. Zhao, S.; Yao, X.; Yang, J.; Jia, G.; Ding, G.; Chua, T.S.; Schuller, B.W.; Keutzer, K. Affective image content analysis: Two decades review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6729–6751. [CrossRef] [PubMed]
22. Wang, Y.; Zhou, S.; Guan, Q.; Fang, F.; Yang, N.; Li, K.; Liu, Y. Enhancing Place Emotion Analysis with Multi-View Emotion Recognition from Geo-Tagged Photos: A Global Tourist Attraction Perspective. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 256. [CrossRef]
23. Ranjan, M.; Tiwari, S.; Sattar, A.M.; Tatkar, N.S. A New Approach for Carrying Out Sentiment Analysis of Social Media Comments Using Natural Language Processing. *Eng. Proc.* **2023**, *59*, 181. [CrossRef]
24. Yang, J.; Li, J.; Li, L.; Wang, X.; Gao, X. A circular-structured representation for visual emotion distribution learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 4237–4246.
25. Yang, J.; Li, J.; Li, L.; Wang, X.; Ding, Y.; Gao, X. Seeking Subjectivity in Visual Emotion Distribution Learning. *IEEE Trans. Image Process.* **2022**, *31*, 5189–5202. [CrossRef] [PubMed]
26. Motiejauskas, M. Convolutional Neural Network Architectures and Training Improvements: Visual Emotion Recognition Case. Ph.D. Thesis, Vilnius University, Vilnius, Lithuania, 2026. [CrossRef]
27. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
28. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
29. Jia, G.; Yang, J. S 2-ver: Semi-supervised visual emotion recognition. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 493–509.
30. Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv* **2016**, arXiv:1610.02136.
31. Mikels, J.A.; Fredrickson, B.L.; Larkin, G.R.; Lindberg, C.M.; Maglio, S.J.; Reuter-Lorenz, P.A. Emotional category data on images from the international affective picture system. *Behav. Res. Methods* **2005**, *37*, 626–630. [CrossRef] [PubMed]
32. Mohammad, S.M.; Kiritchenko, S. An annotated dataset of emotions evoked by art. In Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018), Miyazaki, Japan, 7–12 May 2018.
33. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv* **2016**, arXiv:1608.03983. [CrossRef]
34. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 702–703.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.